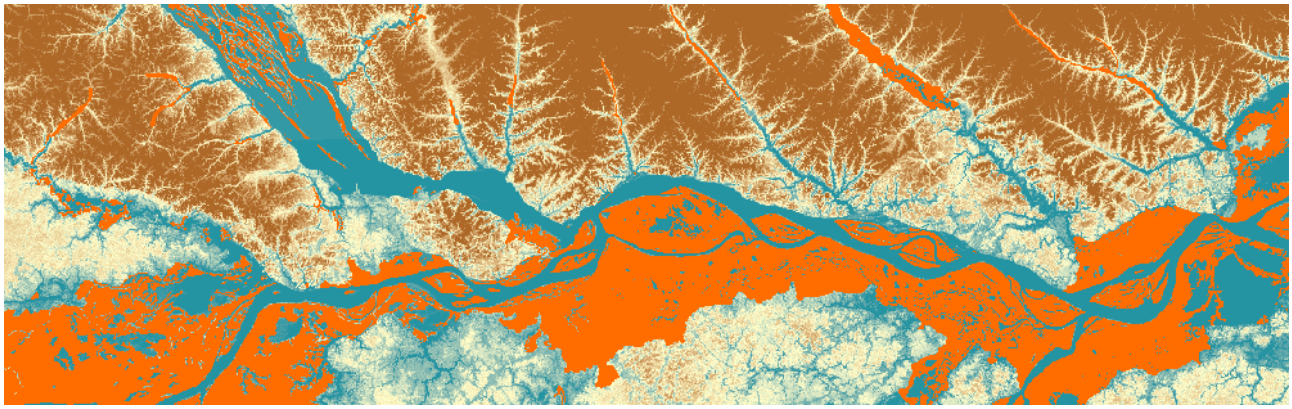
Zürcher Hochschule für Angewandte Wissenschaften
Life Sciences und Facility Management
IUNR Institut für Umwelt und Natürliche Ressourcen
eawag aquatic research

Modelling Groundwater Arsenic Contamination in China with the Groundwater Assessment Platform (GAP)



Bachelor thesis

Arnheiter Ruth

Student of Natural Resource Sciences, BSc

Zurich University of Applied Science, ZHAW, Switzerland

Date: September 7, 2017

Supervisors:

Dr. Joel Podgorski, Project Coordinator GAP, Eawag, Switzerland

Prof. Dr. Baier Urs, Zurich University of Applied Sciences, ZHAW, Switzerland

Picture on front page: gapmaps.org, April 28th, 2017

Abstract

Natural arsenic contamination in groundwater threatens the health of millions of people worldwide. Arsenic prediction models can help policy makers better identify areas of risk. In this study, the GIS-based Groundwater Assessment Platform (GAP, gapmaps.org) was used to produce a prediction model of China, which was compared to the hazard map of Lado et al. published in 2013. Both studies are based on logistic regression using the WHO guideline for arsenic in drinking water of 10 µg L^{-1}, and the same 2668 arsenic measurement data and environmental variables. Lado et al. used eight environmental variables in 100 stepwise logistic regressions that were then aggregated to produce a prediction model. Using a subsample of four of Lado et al.'s variables GAP was used to produce one logistic regression for creating a map of high risk areas. Comparison of the results showed that GAP can produce results as accurate as those of Lado et al. The two maps are highly correlated and show only minor differences when binary coded. A sensitivity analysis showed that the measurement data could be reduced to less than 40% of the initial measurement data and still produces reasonably accurate results, given sufficient variability in the prediction variables used. Some technical limitations and missing information in GAP, such as the number of allowed prediction variables and what criterion is evaluated in the stepwise logistic regression, should be improved to help policymakers better assess the risk of arsenic pollution in groundwater with GAP.

Table of Contents

1 Introduction 6
2 Background 8
2.1 Arsenic in groundwater 8
2.1.1 Sources of arsenic 8
2.1.2 Mobilisation of arsenic 8
2.1.3 Environments related to high arsenic concentration in groundwater 8
2.1.4 WHO guideline of arsenic in drinking water 9
2.2 Predicting arsenic pollution in groundwater with logistic regression 10
2.2.1 Results of logistic regression 12
2.2.2 Evaluation of logistic regression (p-values, coefficients, AUC/ROC) 13
2.2.3 Stepwise logistic regression 13
2.3 Groundwater Assessment Platform (GAP) 14
2.3.1 GAP user interface 14
2.3.2 Modelling process and validation of model in GAP 15
2.4 Arsenic prediction map by Lado et al. (2013) 15
2.4.1 Method of Lado et al. 15
2.4.2 Findings of Lado et al. 16
3 Methods 18
3.1 Modelling process in GAP 18
3.2 Difference analysis between $model_{GAP}$ and $model_{Lado}$ 21
3.3 Sensitivity analysis of GAP 22
4 Results 24
4.1 Probability model created in GAP 24
4.2 Analysed differences between $model_{GAP}$ and $model_{Lado}$ 26
4.3 Analysed sensitivity of GAP 28
5 Discussion 30
6 Conclusion and Outlook 32
Literature 33
7 Index of figures and tables 34
8 Attachments 35
8.1 Article in GapWiki 35
8.2 Signature plagiarism (German) 41

Abbreviations

As	Arsenic
AUC	Area Under the Curve
Fe	Elemental iron
GAP	Groundwater Assessment Platform
GIS	Geographic Information System
O_2	Molecular oxygen
ROC	Receiver Operating Characteristic
TWI	Topographic Wetness Index
WGS	World Geodetic System
WHO	World Health Organisation

1 Introduction

Arsenic (As), occurring as a natural contaminant in groundwater, threatens the health of over 100 million people worldwide (Ravenscroft, Brammer, & Richards, 2009). Since arsenic is tasteless, odourless and colourless, it cannot easily be detected (Eawag, 2015; Holly, 2013). It is toxic already at a few µg L^{-1} (Buschmann et al., 2007) and therefore especially dangerous to human health. In fact, ingestion of arsenic can lead to skin diseases including skin cancer and keratosis (Kinniburgh & Smedley, 2001), especially when exposure occurs over a long period of time (10 - 20 years) (Smith, Lingas, & Rahman, 2000).

High levels of arsenic in groundwater are known in many countries such as Argentina, Chile, China, Hungary, India, Mexico, Bangladesh and Vietnam (Smedley & Kinniburgh, 2002). In 2000, the United Nations defined the eight Millennium Development Goals including Goal 7, Target 10 to halve the number of people without sustainable access to safe drinking water by 2015 compared to 1990 (UN, 2015). In the following years, arsenic levels have been analysed in groundwater in many locations around the world (Berg et al., 2001; Buschmann, Berg, Stengel, & Sampson, 2007; Smedley & Kinniburgh, 2002; Yu, Sun, & Zheng, 2007). However, analysing all water wells worldwide is not feasible on account of available resources.

To overcome this limitation, prediction maps of areas with high arsenic pollutants have been developed to help policymakers identify areas at risk of contamination (Amini et al., 2008; Kinniburgh & Smedley, 2001; Winkel, Berg, Amini, Hug, & Johnson, 2008). In 2001, Kinniburgh et al. (2001) investigated arsenic contamination in Bangladesh using the kriging method, a technique to interpolate measurement data by considering the geospatial context of the sampling site and its relations to the structure of the measurement points (Kinniburgh & Smedley, 2001). In 2008, Winkel et al. predicted groundwater arsenic contamination in southeast Asia based on statistical analysis using logistic regression. The logistic regression calculates the correlation of measurement data with independent (e.g. geospatial) variables (UZH, 2017). For that purpose, they combined measurement data with geological and surface soil variables. In 2013, Lado et al. used logistic regression to create a hazard map of arsenic pollutants in China, basing their model on a wider range of geospatial variables out of the following five categories: geology, surface soil, topography, hydrology and gravity. They defined areas at high and low risk with respect to arsenic pollution by running 100 stepwise logistic regressions. Hence, the technique was not easily applicable, especially for non-professionals.

To make this technique of logistic regression more accessible, the Group Contaminant Hydrology at Eawag designed and developed the free GIS-based online Groundwater Assessment Platform (GAP, gapmaps.org). GAP is an interactive GIS-platform and enables users to create prediction maps by running statistical analysis with logistic regressions within a short time frame and little demand of preparation. With this platform, prediction maps can be produced significantly faster and easier, compared to the previous approaches. GAP was launched in April 2016 but few modelling examples are currently available and so its sensitivity has not yet been determined. More available case studies and a better understanding of the characteristics of the platform are needed to make the promotion of the platform more successful.

This bachelor thesis is therefore focused on the demonstration and explanation of the capabilities of the platform. The aims of the bachelor thesis are (i) to use GAP to produce an arsenic hazard map of China, comparable to that provided by Lado et al., 2013, (ii) to analyse the differences between the two maps, and (iii) to analyse the sensitivity of GAP by determining the number and kind of independent variables needed to produce an effective prediction map.

The results will help Eawag to better understand the limits and capabilities of the platform, which can then aid in making improvements as well as help its promotion and dissemination. It is hoped that GAP will help policymakers to better assess the risk of arsenic pollution in groundwater. As a result, more solutions to reduce arsenic intake by people will hopefully be provided. If so, local people will suffer less from arsenic contamination and hence experience healthier and more productive lives. This would then lead to a better local development and economic stability - one of the three major targets of sustainability (Brown, Hanson, Liverman, & Merideth, 1987).

2 Background

2.1 Arsenic in groundwater

2.1.1 Sources of arsenic

Arsenic, a natural semi-metal, occurs in its elemental form in more than 200 minerals and frequently binds with mineral oxides or iron(hydr)oxides. It is usually incorporated in rocks, especially in sedimentary rocks. Hence, high arsenic concentrations of 3 - 10 mg kg^{-1} are found in sedimentary organic-rich muds and clays. Some of the highest arsenic concentrations (800 – 2900 mg kg^{-1}) are found in Fe-rich ironstone. In addition to the natural sources, industrial pollution from smelting, fossil-fuel combustion products and agricultural products contribute to high arsenic concentrations in soils. Ure and Berrow (1982) found concentrations of up to 732 mg kg^{-1} in orchard soils due to arsenic-containing fungicides and pesticides and exceptionally high concentrations of several thousands of mg kg^{-1} around mining areas. (Boyle & Jonasson, 1973; Smedley & Kinniburgh, 2002; Ure & Berrow, 1982)

2.1.2 Mobilisation of arsenic

Under certain geochemical conditions related to pH, arsenic is released from the solid phase or prevented from adsorbing to the solid phase. These processes occur under oxidising as well as under reducing conditions. Oxidising conditions usually occur in aerobic environments with a high pH. At near-neutral conditions arsenic is strongly bound by oxide minerals. However, the uptake of protons by mineral weathering leads to an increase of pH and arsenic might desorb from mineral oxides, especially when pH increases above 8.5. Such high pH values are often found in arid and semi-arid regions with high rates of evapotranspiration. In contrast, reducing conditions occur in anaerobic environments, as frequently found in alluvial deposits, especially when they are overlaid by a confining layer of finer-grained material. A high level of organic matter in this environment allows for substantial microbial growth, which in turn leads to a relative depletion of O_2. Low levels of O_2 then lead to the decomposition of iron(hydr)oxides and to the release of arsenic. Similar behaviour is also shown in combination with phosphate, bicarbonate and silicate which compete for adsorption sites. (Smedley & Kinniburgh, 2002)

2.1.3 Environments related to high arsenic concentration in groundwater

Previous work has shown that arsenic contamination in groundwater depends on a variety of environmental parameters including geology, soil composition, topography and hydrology (Berg et al., 2001; Buschmann et al., 2007; Manning & Goldberg, 1997; Pei et al., 2010; Ravenscroft, Burgess, Ahmed, Burren, & Perrin, 2005; Rodríguez-Lado et al., 2013; Smedley & Kinniburgh, 2002; Winkel et al., 2008). Ravenscroft (2005), Buschmann (2007) and Winkel et al. (2008)

showed that Holocene depositions are closely linked to high arsenic areas. Holocene sediments developed during the past 10,000 years and hence contain relatively high amounts of arsenic and organic matter, which promote microbial growth with a great demand of O_2 and reducing conditions (Smedley & Kinniburgh, 2002). In Vietnam for example, private wells pump predominantly groundwater from the Holocene aquifers (Berg et al., 2001), making the prediction of areas with high arsenic contamination highly important. The age of soils influences drainage conditions and ion concentrations. Saline soils represent areas with high pH and desorption of arsenic from mineral oxides (Smedley & Kinniburgh, 2002). Older soils such those with significant clay content might absorb arsenic because of their oxide-like characteristic, but the role of binding arsenic is unclear at the present (Smedley & Kinniburgh, 2002; Winkel et al., 2008). The Topographic Wetness Index (TWI) is a proxy for the potential wetness in soils due to topography (Pei et al., 2010). It is a function of the upstream contribution area and the slope (β) of the landscape (Gulens, Champ, & Jackson, 1979) according to the following function:

$$TWI = \ln\left(\frac{A_c}{\tan\beta}\right) \qquad (1)$$

Function of Topographic Wetness Index (TWI). (Rodríguez-Lado et al., 2013)

In which:

A_c = upstream contribution area

B = slop of the landscape at the measurement point

High TWI indicates large catchment areas and/or flat topography at the measurement point. The greater the size of the catchment area, the greater the contribution of arsenic in groundwater. Flat areas might be poorly drained and hence waterlogged. In such flat environments combined with the presence of high organic matter, reducing conditions might occur and arsenic might therefore have more time to accumulate compared to well drained environments (Smedley & Kinniburgh, 2002). Furthermore, small distances to rivers indicate areas with young sediments, which might lead to reducing conditions (Amini et al., 2008). Hence, the density of rivers affects the probability of young sediments and the condition for reducing environments.

2.1.4 WHO guideline of arsenic in drinking water

The World Health Organisation (WHO) recommends the safety guideline of 10 µg L^{-1} arsenic in drinking water. This is a reduction from the previous guideline of 50 µg L^{-1} in 1993 and the guideline value to prevent arsenic carcinogenicity. (WHO, 2011)

2.2 Predicting arsenic pollution in groundwater with logistic regression

As previously mentioned, high arsenic pollution in groundwater can be detected with logistic regression combining measurement points with environmental variables. Logistic regression calculates the correlation of measurements points (dependent variable) with one or more environmental variables (independent variables). It requires (i) the dependent variable to be binary (0/1, no/yes) based on a threshold (e.g. 10 µg L^{-1} arsenic), and (ii) the independent variables not to be highly correlated. The calculation is based on a maximum-likelihood estimation whose function has the shape of an S (Figure 1). The y-values between 0 and 1 are considered to be the probability that the dependent variable will be e.g. 1 or yes in dependency on the independent variables. Hence, the output is not the value of the dependent variable but its probability of being 1 or yes. (UZH, 2017)

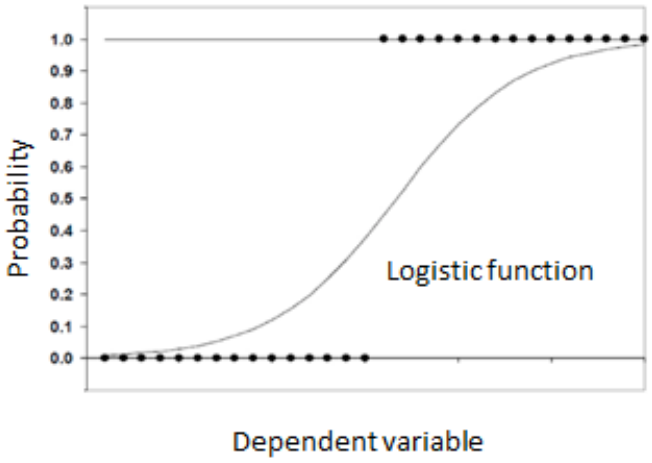


Figure 1 Function of logistic regression. (UZH, 2017)

Logistic regression is calculated with the following function:

$$P(y=1) = \frac{1}{1+e^{-z}} \quad (2)$$

Formula of logistic regression function (UZH, 2017)

In which:
P(y=1) = probability that y=1
e = exponential
z = the “logit” (linear regression model of independent variables)

In which z is calculated as follows:

$$z_k = \beta_0 + \sum_{j=1}^{J} \beta_j \cdot x_{j,k} + u_k \quad (3)$$

The logit used in logistic regression (HSLU, 2017)

In which:
k = case of dependent variable being e.g. 1 or yes
β = coefficient
j = index of independent variable
$x_{j,k}$ = value of the independent variable j in the case defined by k
u_k = error value

2.2.1 Results of logistic regression

The logistic regression creates probabilities between 0 and 1. Each probability represents a trade-off of two functions regarding the true negative (false positive) and true positive (false negative) results (Figure 2). The true negative results (x-axes) are considered to be specificity (ability to correctly classify samples below chosen threshold) and the true positive rates (y-axes) to be sensitivity (ability to correctly classify samples above chosen threshold). As previously mentioned, the probabilities decide only the prediction to be e.g. 1 or yes, and not about the value itself. Consequently, the setting of the probability (called cut-off) determines the prediction to be smaller/equal to or greater than the chosen threshold. For example, a cut-off of 10% predicts almost all of the areas with negative rates (true negative) but little of the areas with positive rates (true positive). In contrast, a cut-off of 100%, predicts most of the area with positive rates but little with negative rates. The crossover marks the cut-off at which point predictions are equal regarding sensitivity and specificity. The most suitable setting of the cut-off depends on the target of the research question. It defines if the cut-off should focus on positive, on negative, or on equally distributed results. (Hosmer, Lemeshow, & Sturdivant, 2013)

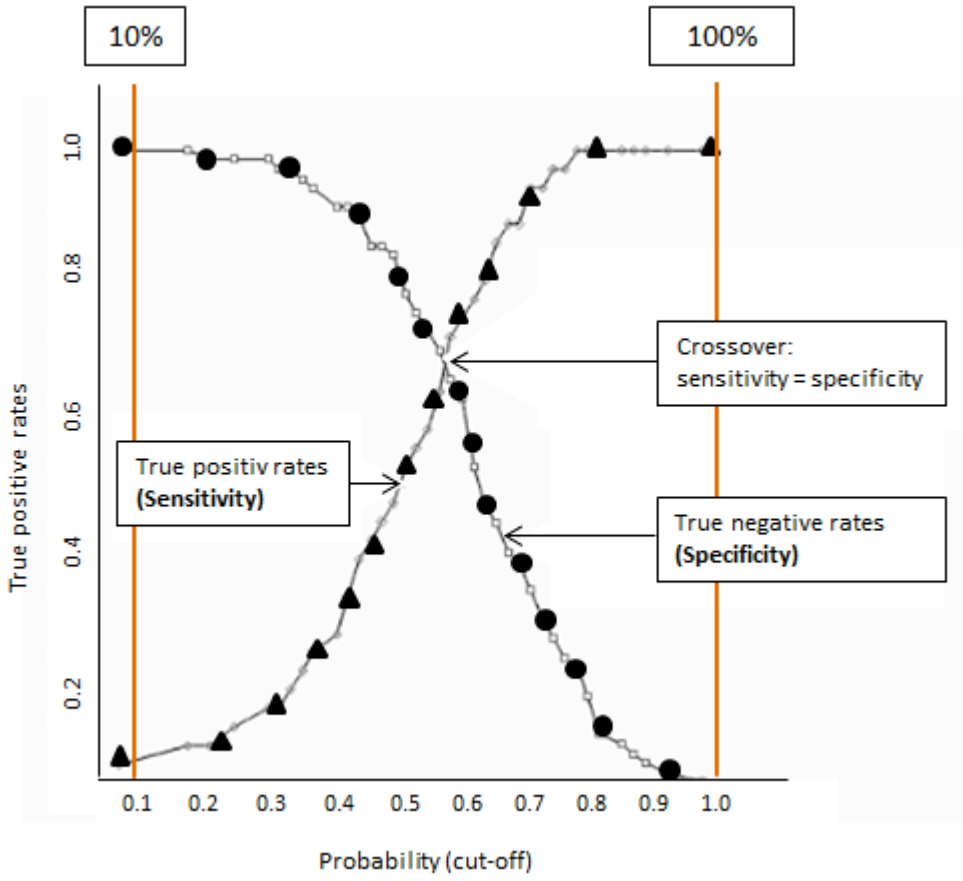


Figure 2 Trade-off of true negative rates (specificity) and true positive rates (sensitivity) of a logistic regression. (Analyse-IT, 2017, modified; GapWiki, 2017a)

2.2.2 Evaluation of logistic regression (p-values, coefficients, AUC/ROC)

Logistic regression calculates the p-value and coefficient of each variable used in the model. The p-value gives information on the statistical significance of the variable; the lower the p-value the higher its significance. The coefficient stands for the power of influence on the model; the greater the number, the more power of the variable on the model. The Receiver Operating Characteristic curve (ROC) plots the previously mentioned probabilities of detecting sensitivity and specificity for all possible cut-offs (Figure 3). The area under this curve (AUC) determines the validity of the model, which defines the model to be good or not. The values range from 0.5 to 1. Values towards 1 represent perfectly run regressions. Values of 0.5 correspond to a random selection. A good model generally requires an AUC of at least 0.7. (Hosmer et al., 2013)

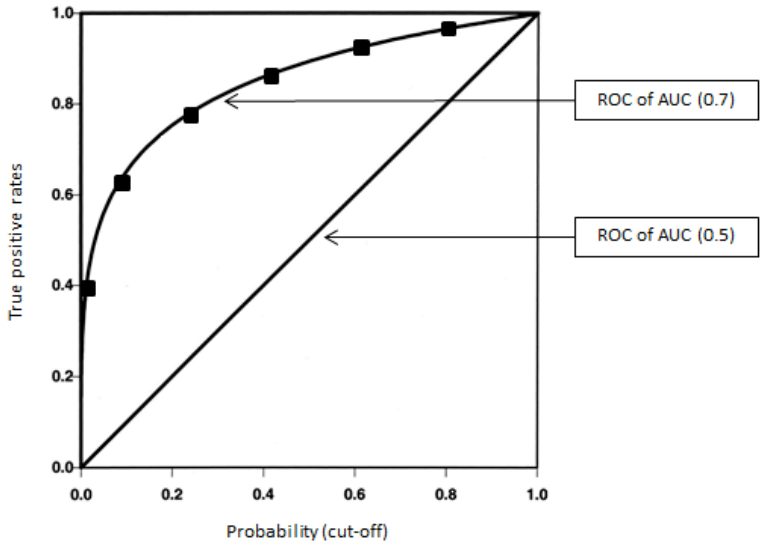


Figure 3 Area under (AUC) the Receiver Operating Characteristic curve (ROC) presenting the probability of detecting true negative and true positive rates for all cut-offs. (Journals.org, 2017, modified)

2.2.3 Stepwise logistic regression

Stepwise logistic regression identifies the independent variables that fit best to the dependent data. It calculates the model multiple times and removes poorly fitting variables at each run. The selection is based on a statistical algorithm (e.g. likelihood ratio) that checks for the importance of each model. Step (0) begins with the calculation of the intercept, an examination of measurement data only, and calculates the corresponding p-value and coefficient. Step (1) joins all variables to one model, evaluates the p-value and coefficient of each variable and compares them with those of the intercept. Variables that do not fit are detected and considered to be the least important ones. All other variables entry step (2) and so on until the best fitting variables are determined. In addition, logistic regression compares each combination of variables on the basis of e.g. log-likelihood of the model by measuring the point where the function of the maximum-likelihood estimation crosses the y-axis. (Hosmer et al., 2013)

2.3 Groundwater Assessment Platform (GAP)

GAP was launched in 2016. It provides GAP Wiki, which contains information on geogenic contamination as well as the possibility to share documents and discuss relevant issues in an open setting. GAP Maps enables users to calculate logistic regression and model maps.

2.3.1 GAP user interface

GAP uses the Google world map as background (Figure 4). By login in, users get the free full function of GAP. Help-link leads the user to GAP Wiki. However, the main features of GAP are stored in the toolbar, where statistical analysis with logistic regression can be run. Map Layers displays the layers to visualise. In addition to the function of uploading measurement data and variables, environmental information such as climate, geology and topography is provided in GAP. Last, but not least, GAP provides additional tools to get pixel values and to change the view. (GapWiki, 2017a)

Figure 4 Main page in GAP. 1 log in, 2 help button which leads to gapmaps.wiki, 3 toolbar, 4 display of all layers user currently works with, 5 provided GAP layers and 6 tools for showing pixel values. (GapWiki, 2017).

2.3.2 Modelling process and validation of model in GAP

GAP requires users to upload desired measurement data and environmental variables if desired. Users define the threshold the logistic regression should be based on, combine the measurement data with the environmental variables und run the logistic regression (normal or stepwise) to model the prediction map. To validate the model GAP provides p-values and coefficients for each variable and information on sensitivity and specificity regarding ROC curve as well as the AUC. (GapWiki, 2017b; Hosmer et al., 2013)

2.4 Arsenic prediction map by Lado et al. (2013)

In China, arsenic's impact on the health of the population was first recognized in the 1960's. Despite the fact that well testing was done by the Chinese National Survey Program between 2001 and 2005, 95% of all wells countrywide remained untested (Yu et al., 2007). In 2013, Lado et al. developed a prediction model based on logistic regression calculations to predict areas with high arsenic probabilities throughout China. They cross-referenced their results with a population map and estimated that 19.6 million people countrywide are exposed to high-risk groundwater. (Rodríguez-Lado et al., 2013)

2.4.1 Method of Lado et al.

Based on the arsenic measurements of the Chinese National Survey Program, Lado et al. used 2,668 measurement data throughout China. Out of five categories, they used eight environmental variables (Holocene sediments, saline soils, subsoil texture, TWI, slope, density of rivers, distance to rivers and gravity). The variables were all retrieved from open source databases. Saline soils included Solonchaks, Solonetzs and soils with a salic phase, which all indicate soils with high salinity. Subsoil texture was grouped by clay/loam, sandy/clay and loam, which are all clayey soils. All variables, except subsoil texture, were significant (p-value: >0.05) at the univariate logistic regression, which tests a variable's ability to explain the measurement data. However, subsoil texture was included due to the fact that fine and medium soil textures clearly affect the arsenic concentration in groundwater. In addition, all variables were normalized to the same range of values and aggregated to 1 km^2 resolution before modelling. Lado et al. ran 100 stepwise logistic regressions, each with a random subset of measurement points (80% of 2668 measurement data) by sampling with replacement. They used the WHO guideline for arsenic contamination in drinking water (10 µg L^{-1}) and grouped together all best fitting variables using the Akaike Information Criterion (AIC). AIC uses the function of the likelihood and the number of variables of each model to compare models with different numbers of parameter (Hosmer et al., 2013). In addition, they calculated the overall internal accuracy ($p > 0.05$) of each group (model) with the Hosmer & Lemeshow goodness-of-fit test, which estimates how well the model fits the probability predictions

(Hosmer et al., 2013). 48 models passed this test and were used to build the final model. For each variable used in the final model, its frequency in the 48 models and the corresponding coefficient were calculated to predict the power of influence of each variable. The final model was categorized in areas with low or high probabilities with the cut-off value of 0.46, estimated using the Receiver Operating Characteristic curve (ROC), which plots probabilities of detecting sensitivity and specificity for all possible cut-offs (Hosmer et al., 2013). They validated the model with another group of 625 arsenic measurement data composed of 184 measurement data from the Chinese National Survey Program, 261 published measurement data and 180 measurement data from their own field surveys. The model accuracy was then estimated from the overall correctly classified rate of sensitivity and specificity. (Rodríguez-Lado et al., 2013)

2.4.2 Findings of Lado et al.

Holocene sediments, soil salinity, soil texture and TWI were the most important variables for the final model by showing frequencies between 45 and 48 (Table 1). By comparison, slope, density of rivers, distance to rivers and gravity showed smaller frequencies (between 1 and 4) and were considered to be less important to their final model. Out of the four most frequent variables the highest coefficient was shown by TWI (4.1), followed by saline soils (1.04), Holocene sediments (0.91) and subsoil texture (0.16). The probability model showed high risk areas in the Xinjiang province (1), Hetao-Huhhot basin (2), Minqin basin and Chahaertan oasis (3), as well as in the Liao-Ho basin (4) (Figure 5). The former two provinces are characterized by rather anoxic environments with flat topography and high salinity. The last two areas contain oxic aquifers showing oxidising mobilisation. (Rodríguez-Lado et al., 2013)

Table 1 Summary of the independent variables used in Lado et al.'s survey and frequencies and coefficients of each variable retrieved true 100 stepwise logistic regressions. (Rodríguez-Lado et al., 2013)

Category	**Variable**	*Frequency*	*Coefficient*
Geology	Holocene sediments	48	0.91
Soils	Saline Soils	48	1.04
	Subsoil texture	47	0.16
Topography	TWI	45	4.1
	Slope	4	-17.21
Hydrology	Density of rivers	7	1.02
	Distance to rivers	3	-11.53
Gravity	Gravity	1	-4.72

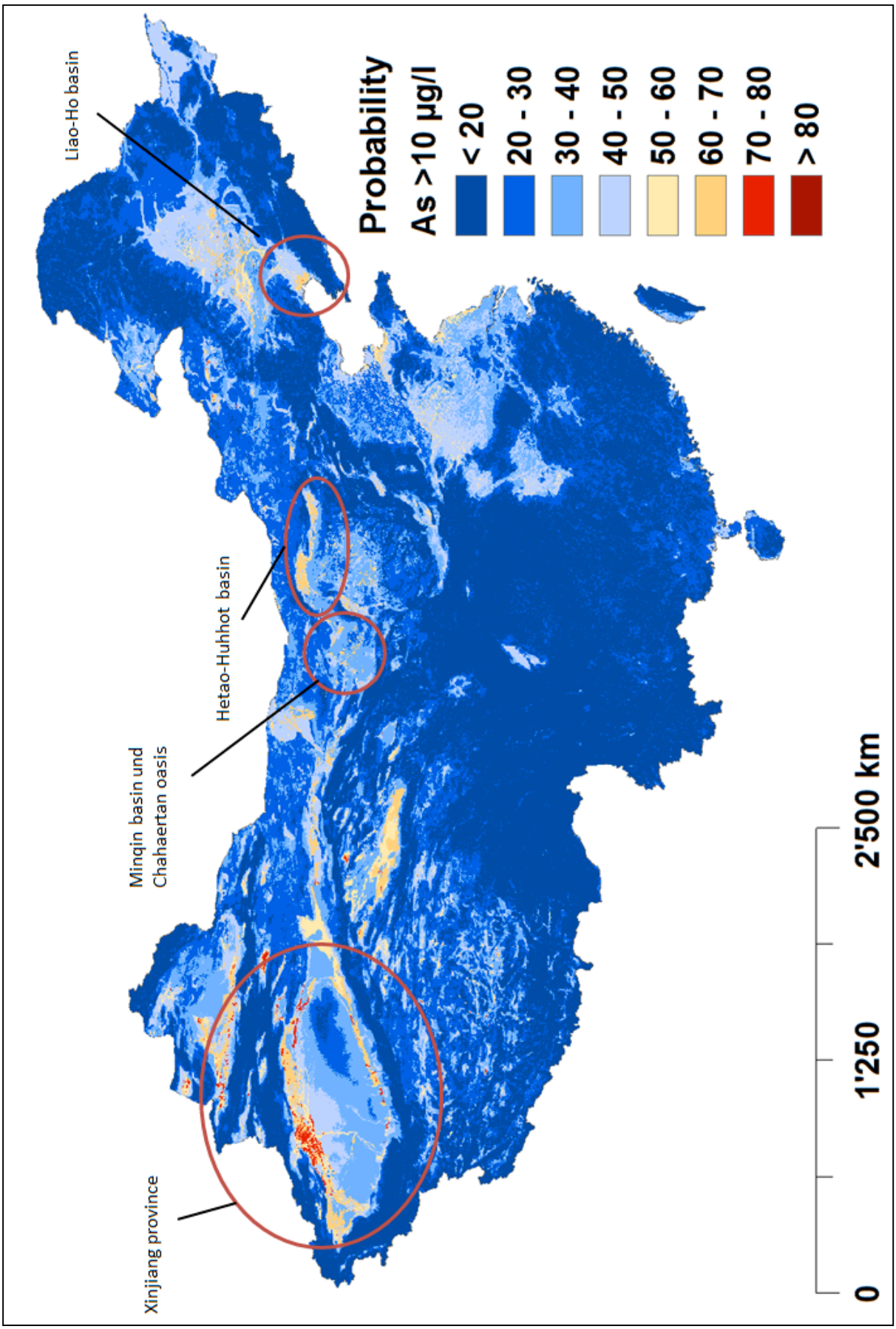


Figure 5 Prediction map of arsenic probability greater than 10 µg/l in China produced by Lado et al. Rectangles indicate area with high risk probabilities. (Rodríguez-Lado et al., 2013, modified)

3 Methods

3.1 Modelling process in GAP

Based on Lado et al.'s survey, 2668 arsenic measurements and four environmental variables (Holocene sediments, saline soils, subsoil texture and TWI) which appeared with the highest frequencies in Lado et al.'s study, were assembled for modelling (Table 2, Figure 6 and 7). Lado et al.'s other four used variables (slope, density of rivers, distance to rivers and gravity) could not be included because of technical requirements of GAP. In this study, GAP allowed only six variables to run the stepwise logistic regression. The reasons for the limitation were not clear since GAP usually does not restrict the number of variables. The variables were dismissed according to the following rules: (i) slope was dismissed because it correlated highly with TWI, and (ii) distance to rivers was dismissed to establish an equal distribution of all five categories and due to the fact that it showed the lowest frequency in its category. Gravity and density of rivers were later dismissed during the preparation phase with stepwise logistic regression. Subsequently, logistic regression was calculated with the WHO guideline for arsenic in drinking water of 10 µg L^{-1} and run once. The generated prediction model ($model_{GAP}$) was transferred to ArcGIS.

	GAP	*Lado et al.*
Measurement data	2668	2668
Validation data	-	625
Number of Variables	4	8
Variables	Holocene sediments Saline soils Subsoil texture TWI	Holocene sediments Saline soils Subsoil texture TWI Slope Density of rivers Distance to rivers Gravity
Modelling	1 stepwise log. regression 1 logistic regression	100 stepwise log. regressions

Table 2 Summary of measurement data, validation data, number and kind of variables and modelling approach in GAP and in Lado et al.'s study.

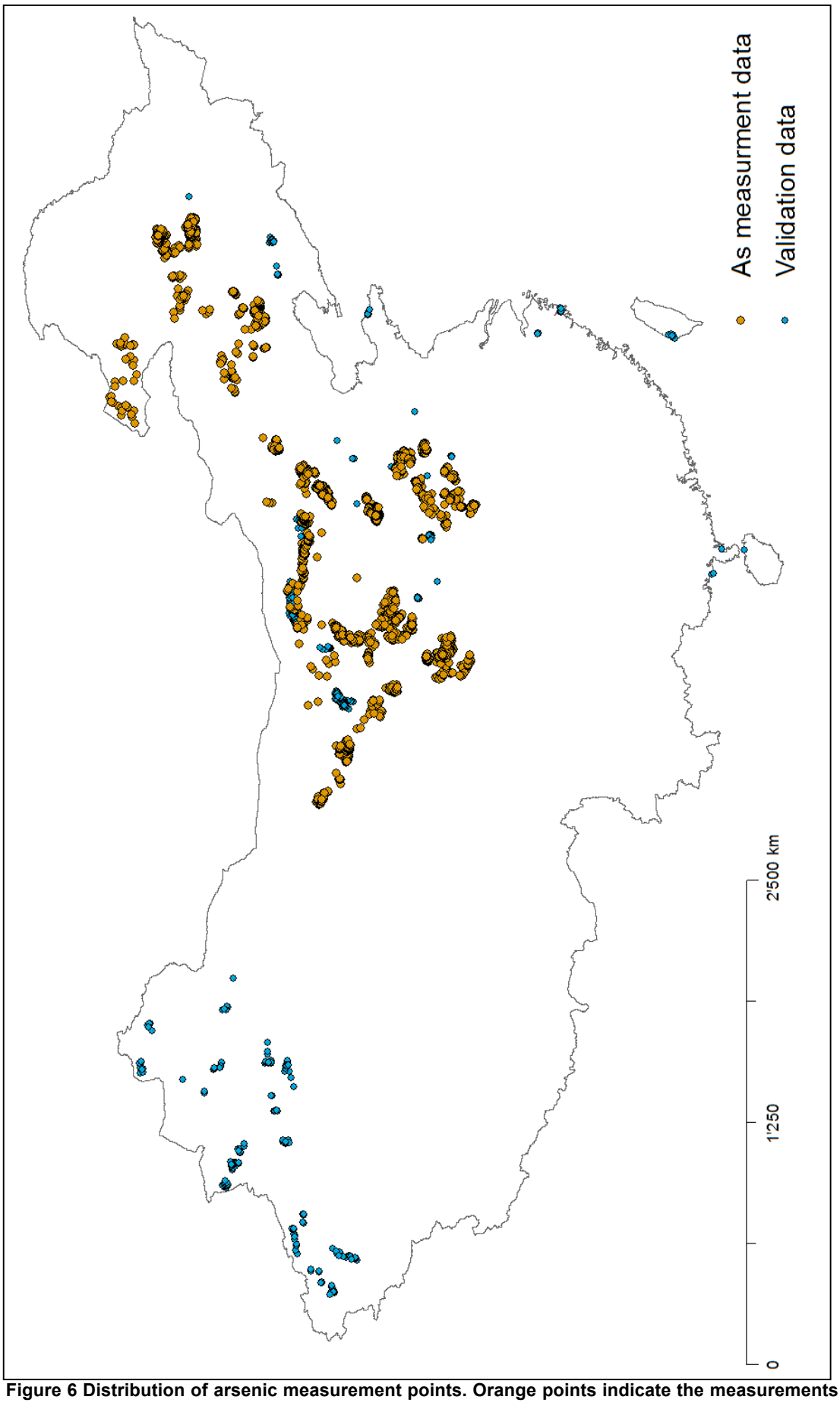


Figure 6 Distribution of arsenic measurement points. Orange points indicate the measurements data Lado et al. and this study based the calculations on, blue points are the measurement data Lado et al. used for the validation of the model.

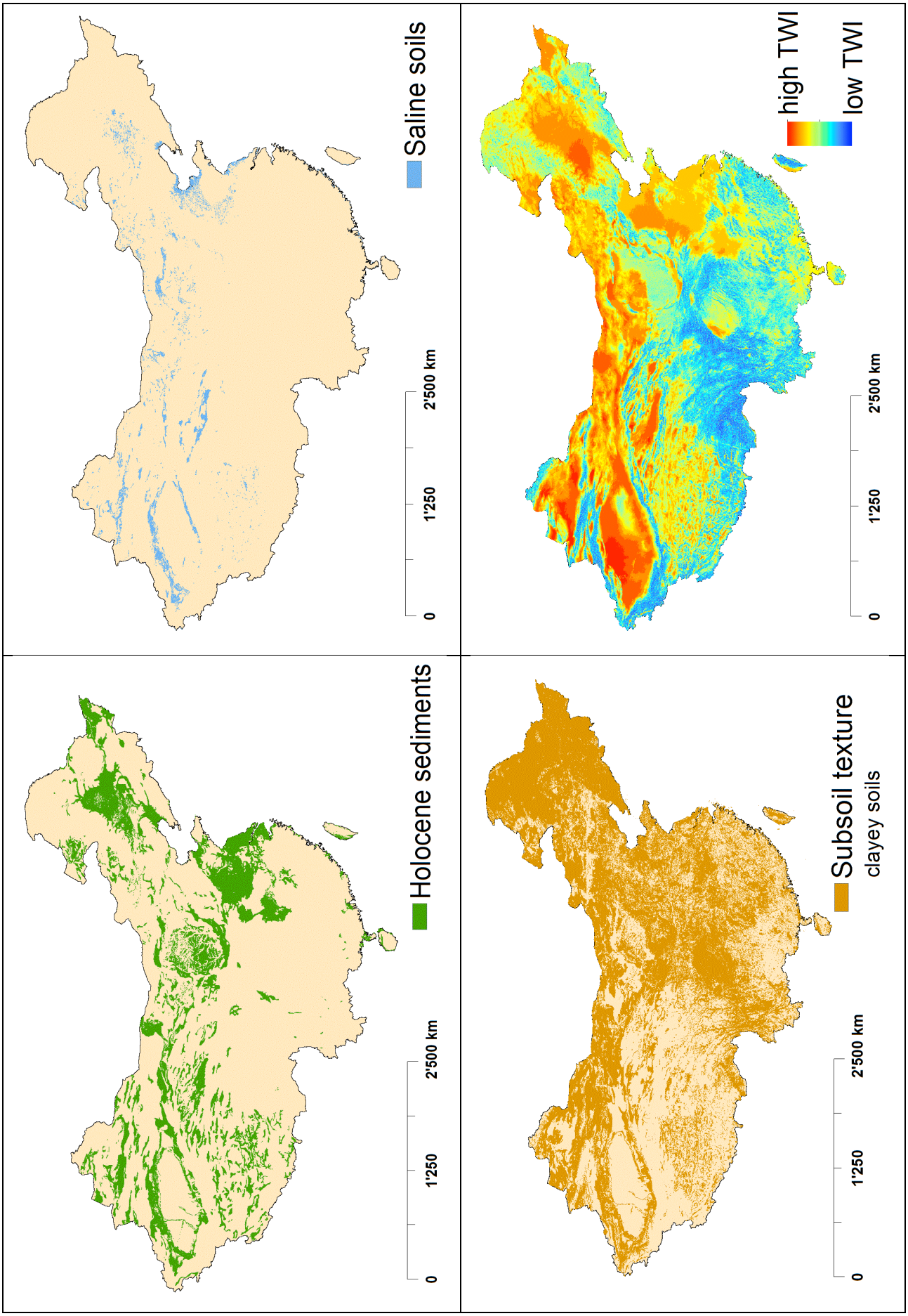


Figure 7 Auxiliary variables included in this study.

3.2 Difference analysis between $model_{GAP}$ and $model_{Lado}$

In ArcGIS (version 10.4), $model_{GAP}$ was aligned to the same grid with Lado et al.'s hazard map ($model_{Lado}$) for accurate comparison. The correlation of the probabilities of $model_{GAP}$ and $model_{Lado}$ were calculated with the band collection statistic tool in ArcGIS.

In addition, the models were binary coded using the cut-off values of 0.41 for $model_{GAP}$ and 0.46 for $model_{Lado}$. The cut-off for $model_{GAP}$ was compiled in GAP by evaluating the trade-off of sensitivity and specificity (Figure 8). The cut-off of 0.46 was determined by Lado et al. using the ROC curve as previously mentioned (Rodríguez-Lado et al., 2013). Subsequently, the differences of the two binary models were located by subtracting one model from the other and the differences in percentages calculated. In addition, the localised differences were compared with the probability map to better understand the relationship between the differences and predicted probabilities. To estimate the influence of the cut-off $model_{GAP}$ was again binary coded with the cut-off of 0.46 and processed as previously.

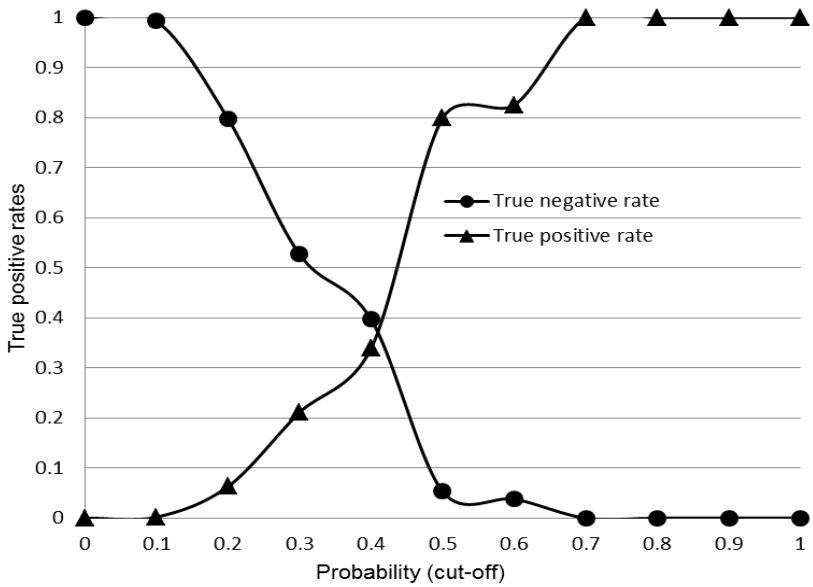


Figure 8 Trade-off of true negative and true positive rates of the logistic regression for $model_{GAP}$ revealed in GAP.

3.3 Sensitivity analysis of GAP

Once a suitable model using all 2668 measurement data was created, the effect of using fewer measurement data was tested in order to better understand the sensitivity of GAP. Randomly selected subsets of the 2668 measurement data were retrieved (250, 500, 750, 1000, 1250, 1500 and 2000 measurement data) in threefold repetition (Table 3). In addition, a data set of 3318 measurement data was aggregated from 2668 measurement data and 625 validation data used in Lado et al.'s study. There was an addition of 25 measurement data which might have been generated by the previous aggregation of the measurement points to 1 km^2 resolution. Each data set was processed according to the previous calculations with stepwise logistic regression to calculate the relative and binary correlations using the corresponding cut-offs compiled in GAP. In addition, for a better understanding of the vulnerability of the correlations, the model with 1250 measurement data was recalculated in tenfold repetition. To examine the effect of using fewer variables on the modelling, the validity of all combinations of the four variables used in $Model_{GAP}$ (Holocene sediments, saline soils, subsoil texture, TWI) were analysed by comparing the corresponding AUC (Table 4). Every combination of variables was joined with the 2668 measurement data to run a logistic regression. To prove that non-related variables produce an AUC around 0.5, the 2668 measurement data were joined with a random variable temperature to calculate the corresponding AUC.

Table 3 Summary of all data sets, repetitions and corresponding cut-offs used for the sensitivity analysis of GAP. The data set of 1250 measurement data was run in tenfold repetition to better understand the vulnerability of the correlations.

Number of measurement data	*Repetitions*	*Cut-off*
250	3	0.34, 0.4 and 0.41
500	3	0.32, 0.34 and 0.42
750	3	0.33, 0.37 and 0.41
1000	3	0.42, 0.43 and 0.47
1250	3/10	0.38, 0.41 and 0.42 (threefold repetition) / 0.38, 0.40, 0.40, 0.41, 0.41, 0.42, 0.43, 0.43, 0.44 and 0.44 (tenfold repetition)
1500	3	0.42, 0.42, 0.42
2000	3	0.41, 0.41, 0.41
3318	1	0.47

Table 4 Summary of all analysed combinations out of the four variables (Holocene sediments, saline soils, subsoil texture, TWI) to detect the corresponding AUC for the examination of needed variables.

Holocene sediments		x				x	x	x				x	x	x		x
Subsoil texture			x			x			x	x		x	x		x	x
Saline soils				x			x		x		x	x		x	x	x
TWI					x			x		x	x		x	x	x	x
Temperature	x															
Number of variables	1	1	1	1	1	2	2	2	2	2	2	3	3	3	3	4

4 Results

4.1 Probability model created in GAP

In $model_{GAP}$, all variables showed p-values < 0.01, except for soil texture, which presented a p-value of 0.03 (Table 5). The highest coefficient were shown by saline soils (0.88), followed by Holocene sediments (0.73), soil texture (-0.25) and TWI (0.18). In addition, soil texture was showed a negative coefficient. The AUC of the model was 0.68 (Figure 9). $Model_{GAP}$ showed probabilities between 20 - 80% (Figure 10). High risk areas were determined in the Xinjiang and Quinghai province, and the Hetao-Huhhot and Liao-Ho basins (Figure 10).

Table 5 P-values of all variables used calculated by GAP.

Variable	*p-value*	*coefficients*
Holocene sediments	< 0.01	0.73
Saline soils	< 0.01	0.88
Soil texture (clayey soils)	0.031	-0.25
TWI	< 0.01	0.18

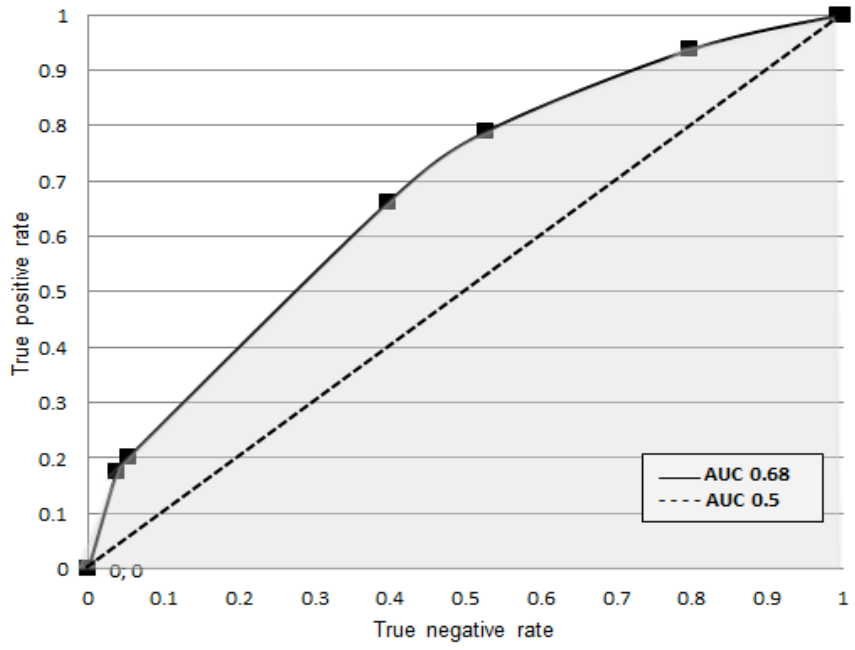


Figure 9 Receiver Operating Characteristic Curve (ROC) of modelGAP and the corresponding area under the curve (AUC) retrieved in GAP

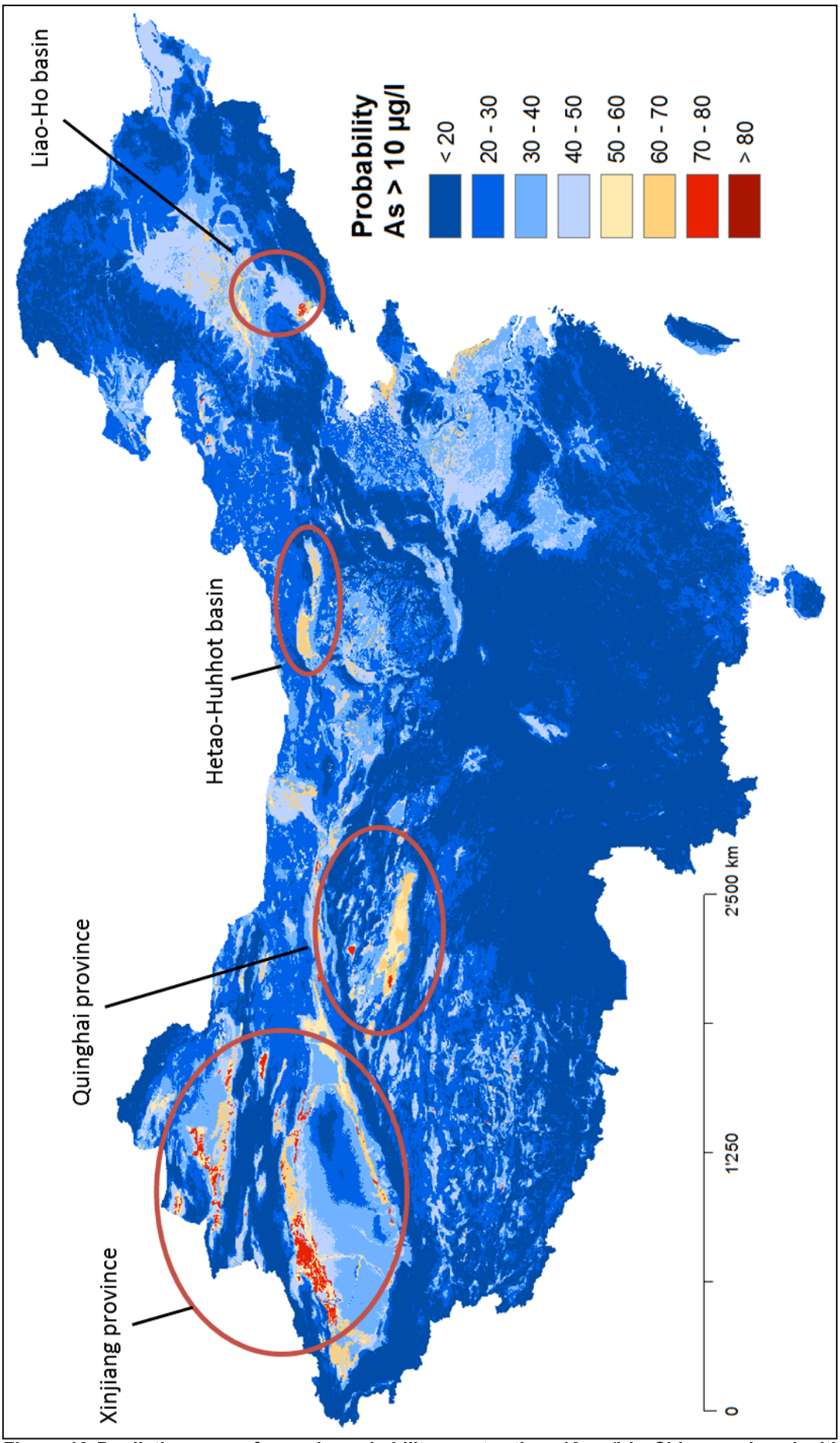


Figure 10 Prediction map of arsenic probability greater than 10 µg/l in China produced with GAP. Circles indicate areas with high risk probabilities.

4.2 Analysed differences between $model_{GAP}$ and $model_{Lado}$

The probabilities of $model_{GAP}$ and $model_{Lado}$ were highly correlated (0.97) (Table 6). The binary models using the cut-off 0.41 for $model_{GAP}$ differed in 6% of the area (Figure 11). 95% of these differences lied in areas with predicted probabilities of high arsenic contamination in groundwater between 40 – 50%. Using the cut-off of 0.46 for $model_{GAP}$, the two binary $model_{GAP}$ and $model_{Lado}$ were brought more in line. The differences became smaller, in fact as low as 1.5%.

Table 6 Correlation between the probability $model_{GAP}$ and $model_{Lado}$ and congruence (%) between the binary models with the cut-off 0.41 and 0.46 for $model_{GAP}$.

	Correlation	***Congruence (%)***	***Congruence (%)***
	$Model_{GAP}$ (probability)	$Model_{GAP}$ (binary; cut-off 0.41)	$Model_{GAP}$ (binary; cut-off 0.46)
$Model_{Lado}$ (probability)	0.97	-	-
$Model_{Lado}$ (binary; cut-off 0.46)	-	6	1.5

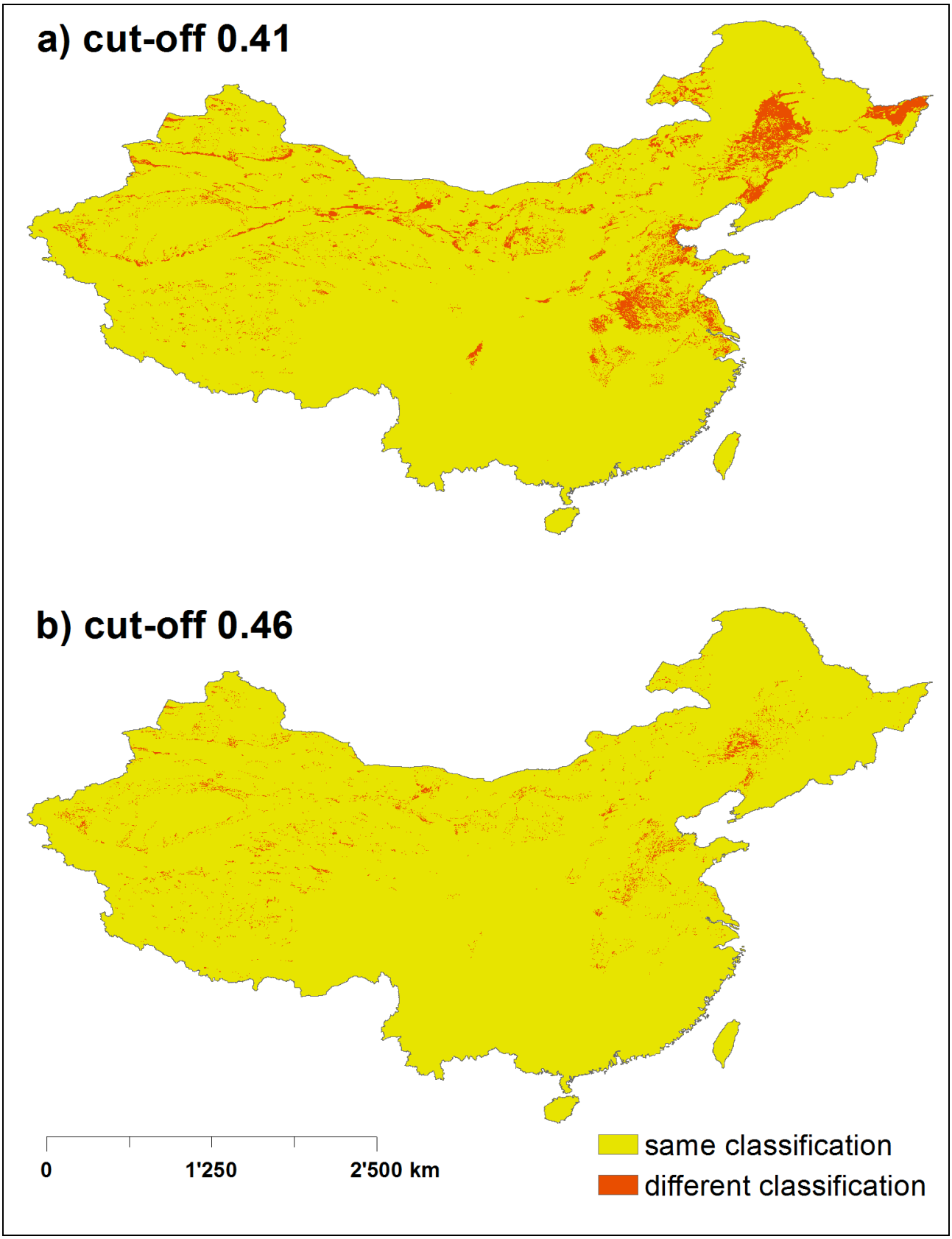


Figure 11 Localised differences between the binary model$_{GAP}$ and model$_{Lado}$ (a) with the cut-off 0.41, and b) with the cut-off 0.46.

4.3 Analysed sensitivity of GAP

As expected, the best result was achieved with 2668 measurement data (Figure 12). Most of the randomly selected subsets above the number of 1000 measurement data showed similar results compared to the initial 2668 measurement data. The probability models correlated well (above or around 0.95), the binary models were congruent at approximately 90% and the AUC was no smaller than 0.67 for any data set. Data sets lower than 1000 measurement data achieved less reliable results, considering means, variations and AUCs. Even the data set of 1250 measurement data still showed great variations, especially with the binary models. This variation, however, was reduced when the calculations were repeated multiple times (Figure 13). Testing the required number of variables proved that the best result was achieved with the initial four variables Holocene sediments, saline soils, subsoil texture, and TWI (0.68) (Figure 14). The smallest AUC was shown with the variable subsoil texture (0.5). An AUC of 0.65 was achieved by using either three variables (Holocene sediments, subsoil texture and saline soils) or two variables (Holocene sediments and TWI or Holocene sediments and saline soils). Temperature showed an AUC of 0.51, indicating that temperature does not contribute to arsenic contamination.

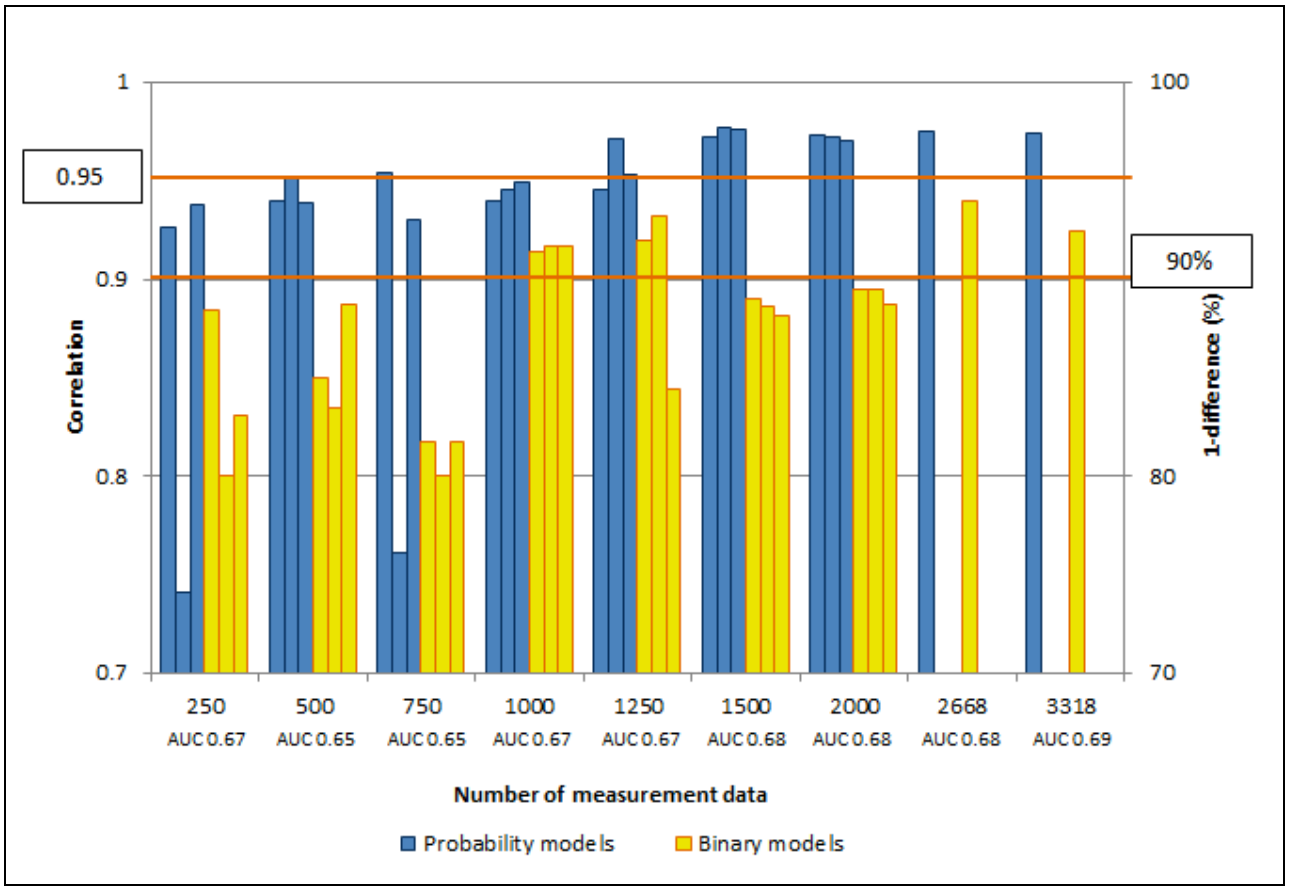


Figure 12 Correlations of probability $model_{GAP}$ and $model_{Lado}$ (blue) and 1-difference (%) of the binary models (yellow) as well as the corresponding AUC of different data sets of 250, 500, 750, 1000, 1250, 1500 and 2000 measurement data in threefold repetition and of the original data set of 2668 measurement data as well as the data set of 3318 measurement data with one repetition.

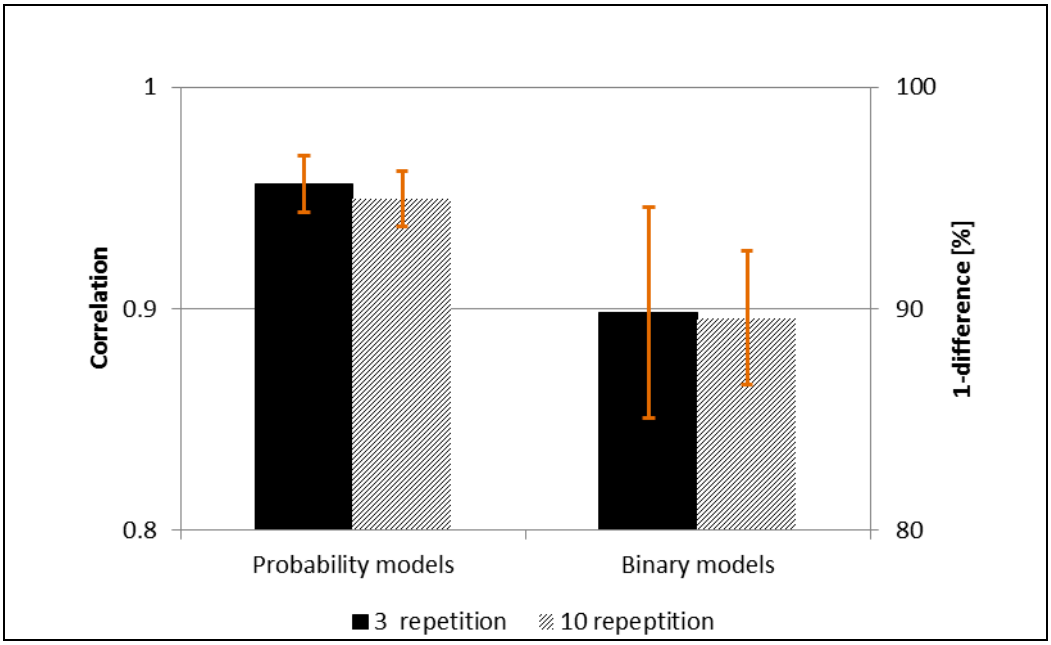
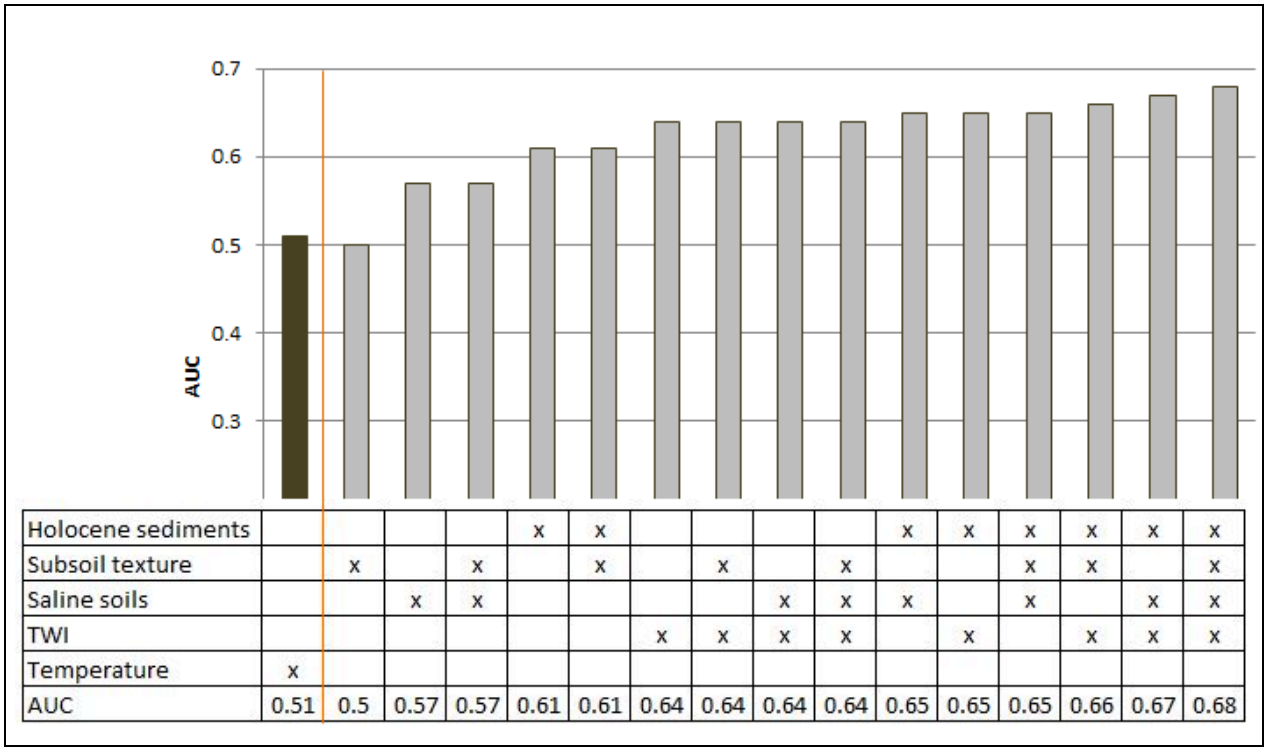


Figure 13 Correlations and standard deviations of probability model$_{GAP}$ and model$_{Lado}$ and binary models with 3 and 10 repetitions.



Holocene sediments					x	x					x	x	x	x	x	x
Subsoil texture		x		x		x		x		x			x	x		x
Saline soils			x	x					x	x	x		x		x	x
TWI							x	x	x	x		x		x	x	x
Temperature	x															
AUC	0.51	0.5	0.57	0.57	0.61	0.61	0.64	0.64	0.64	0.64	0.65	0.65	0.65	0.66	0.67	0.68

Figure 14 Summary of AUCs of all combinations of the four variables Holocene sediments, saline soils, subsoil texture, TWI used for model$_{GAP}$.

5 Discussion

GAP predicted the same areas to be high risk areas as Lado et al. Holocene sediments, saline soils, subsoil texture and TWI are indeed the more important variables determining arsenic contamination in groundwater, compared to slope, density of rivers, distance to rivers and gravity. The negative correlation of subsoil texture showed that the presence of clayey soils decreases the probability of arsenic in groundwater. The variables used in the models, though, differed in terms of their influences on the model. Lado et al. predicted the variable TWI to be the most influencing one among all variables used but GAP defined it to be the least influencing one from among the four variables used for this study. In addition, the best achieved AUC in GAP was 0.68, which is considered to be too low for a good prediction model (Hosmer et al., 2013). Nevertheless, GAP produces results as accurate as those of Lado et al. by running only one logistic regression, compared to the 100 stepwise logistic regressions that Lado et al. used. In this survey, some technical requirements of GAP limited the number of variables with which the stepwise logistic regression could be run. This limitation was most likely due to a bug in the programming language. GAP, however, did not give further information on how users should proceed. Without having had Lado et al.'s preparatory study as a guideline, a wrong collection of variables could have been selected due to the technical limitations in GAP and the results may have differed. In the future, GAP should allow users to add any number of variables to prevent users from making assumptions about which variables are important. Furthermore, it is not clear on which information GAP defines a variable to fit or not to fit. It might be possible that GAP bases the stepwise logistic regression on p-values as defined by Hosmer (2013) but it is also possible that GAP uses the Akaike Information Criterion (AIC). AIC uses the function of the likelihood and the number of variables of each model to compare models with different numbers of parameters (Hosmer et al., 2013). More information on how GAP decides a variable to fit or not to fit could be helpful for users to better understand the techniques behind GAP.

Despite the different number of variables (four variables for $model_{GAP}$, eight variables for $model_{Lado}$), the two models' probabilities were highly correlated (0.97). The binary model with the cut-off of 0.41 slightly differed from the binary model of Lado et al (6%). Interestingly, almost all localised differences were localised in areas with risk probabilities between 40 - 50% (95%). Approximating the cut-off for $model_{GAP}$ to that of $model_{Lado}$, the area of differences decreased to one fourth of the differences revealed with the smaller cut-off (1.5%). Hence, the threshold used for the binary coding is decisive about the correlation of the two binary models. Nevertheless, whatever technique was used for the analysis of differences, the two models showed a congruency of 94%, which is very high. It is not clear which exact calculations Lado et al. used to set the cut-off by using the ROC curve. In contrast, for each logistic regression GAP provides a figure with the

trade-off of true positive and false positive in which the cut-off can be read out. This figure helps users better select the corresponding cut-off. Mentioning the exact number of the cut-off (for example by pop-up windows while moving the mouse over the figure) will even be more easily understandable and precise.

Most of the randomly selected subsets of the 2668 measurement data showed a similar behaviour compared to the initial 2668 measurement data. As expected, the bigger the data set was, the better the models were correlated. Interestingly, similar results (compared to the 2668 measurement data) were achieved with as little as 1000 measurement data, which is less than 40% of the initial data set. However, these investigations were based on only three repetitions, which may not be enough to be conclusive, especial when standard deviations vary, such as is the case for the set of 1250 measurement data. Nevertheless, the investigations with ten repetitions showed that the higher the number of repetitions, the smaller the variation. The investigation of the variables showed the best results with four of them. Interestingly, in some cases, the number of variables did not determine the AUC; rather the composition of the variables did. However, all combinations of the four variables produced a smaller AUC than each of the four variables by themselves. Temperature as a variable showed that an AUC of 0.51 indicates non-correlation. Nevertheless, this study does not give final recommendations regarding the number of required measurement points related in relation to the area of investigation. For example, downscaling these results to Switzerland, which has a surface area of 41,285 km^2 (which is less than 0.5 % of China's 9,297,000 km^2), would mean that less than 0.5% of 1000 measurement data should be sufficient to predict arsenic pollution in Switzerland, which intuitively is insufficient for a downscaled area such of Switzerland. In the future, the determination of the lower limit of measurement data sets in relationship to the surface area still producing reliable results will help policymakers to better determine how many measurements points are required. In addition, this survey did not analyse the potential influence of the local distribution of measurement points. Investigating whether such distribution is indeed of importance for the generation of prediction maps should be of considerable interest for further studies.

6 Conclusion and Outlook

GAP produces results as accurate as those of Lado et al. by running one logistic regression, compared to Lado et al. using 100 stepwise logistic regressions. The measurement data could be limited to 1000 data, as well as to four variables in this survey. The selection of the most important variables was essential, though. Therefore, in the future, GAP should be configured to (i) allow users to add any number of variables and (ii) give information on how the program defines a variable to fit or not to fit. These improvements of GAP will help users to make the best selection of variables and to better understand the techniques behind the platform and thereby enhancing GAP's promotion and dissemination.

The following improvements of GAP and further investigation could be of interest:

- In which cases do the technical limitations of GAP appear? Is it possible to overcome these technical limitations?
  - *This will prevent users from making assumptions about which variables are important.*

- How does GAP define a variable to fit or not to fit?
  - *This will help users to better understand the techniques behind GAP.*

- Could the output figure of sensitivity and specificity in GAP be added with the exact cut-off (e.g. by pop-up windows?)
  - *This will help users to better set the cut-off.*

- Which is the smallest measurement data set related to the area of interest that still produces reliable results in GAP?
  - *This will help policymakers to determine the number of measurements points required.*

- How does the distribution of measurement data affect the probability modelling in GAP?
  - *This will help policymakers to determine the distribution of measurements points required.*

Literature

Amini, M., Abbaspour, K. C., Berg, M., Winkel, L., Hug, S. J., Hoehn, E., & Johnson, C. A. (2008). Statistical modeling of global geogenic arsenic contamination in groundwater. *Environmental science & technology, 42*(10), 3669-3675.

Analyse-IT. (2017). Method evaluation of ROC curve plot.

Berg, M., Tran, H. C., Nguyen, T. C., Pham, H. V., Schertenleib, R., & Giger, W. (2001). Arsenic contamination of groundwater and drinking water in Vietnam: a human health threat. *Environ Sci Technol, 35*(13), 2621-2626.

Boyle, R., & Jonasson, I. R. (1973). The geochemistry of arsenic and its use as an indicator element in geochemical prospecting. *Journal of Geochemical Exploration, 2*(3), 251-296.

Brown, B. J., Hanson, M. E., Liverman, D. M., & Merideth, R. W. (1987). Global sustainability: toward definition. *Environmental management, 11*(6), 713-719.

Buschmann, J., Berg, M., Stengel, C., & Sampson, M. L. (2007). Arsenic and manganese contamination of drinking water resources in Cambodia: coincidence of risk areas with low relief topography. *Environ Sci Technol, 41*(7), 2146-2152.

Eawag. (2015). Geogenic Contamination Handbook - Addressing Arsenic and Fluoride in Drinking Water. C.A. Johnson, A. Bretzler (Eds.),

: Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dübendorf, Switzerland.

GapWiki. (2017a). GAP Help. Retrieved from http://www.gapmaps.wiki/index.php?title=GAP_Help

GapWiki. (2017b). Statistical Analysis. Retrieved from http://www.gapmaps.wiki/index.php?title=Statistical_Analysis

Gulens, J., Champ, D., & Jackson, R. (1979). Influence of redox environments on the mobility of arsenic in ground water: ACS Publications.

Holly, M. (2013). An arsenic forecast for China. *Science, 341*(6148), 852-853.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression, Third Edition*: Willey, New Jersey.

HSLU. (2017). Hochschule Luzern: Logistic regression empirical methods Retrieved from https://www.empirical-methods.hslu.ch/entscheidbaum/zusammenhaenge/logistische-regression/

Journals.org. (2017). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predicitve Models. Retrieved from http://circ.ahajournals.org/content/115/5/654

Kinniburgh, D., & Smedley, P. (2001). Arsenic contamination of groundwater in Bangladesh.

Manning, B. A., & Goldberg, S. (1997). Adsorption and stability of arsenic (III) at the clay mineral− water interface. *Environmental science & technology, 31*(7), 2005-2011.

Pei, T., Qin, C. Z., Zhu, A. X., Yang, L., Luo, M., Li, B. L., & Zhou, C. H. (2010). Mapping soil organic matter using the topographic wetness index: A comparative study based on different flow-direction algorithms and kriging methods. *Ecological Indicators, 10*(3), 610-619. doi:10.1016/j.ecolind.2009.10.005

Ravenscroft, P., Brammer, H., & Richards, K. (2009). *Arsenic pollution: a global synthesis* (Vol. 28): John Wiley & Sons.

Ravenscroft, P., Burgess, W. G., Ahmed, K. M., Burren, M., & Perrin, J. (2005). Arsenic in groundwater of the Bengal Basin, Bangladesh: Distribution, field relations, and hydrogeological setting. *Hydrogeology Journal, 13*(5-6), 727-751.

Rodríguez-Lado, L., Sun, G., Berg, M., Zhang, Q., Xue, H., Zheng, Q., & Johnson, C. A. (2013). Groundwater arsenic contamination throughout China. *Science, 341*(6148), 866-868.

Smedley, P. L., & Kinniburgh, D. G. (2002). A review of the source, behaviour and distribution of arsenic in natural waters. *Applied Geochemistry, 17*(5), 517-568. doi:Pii S0883-2927(02)00018-5

Doi 10.1016/S0883-2927(02)00018-5

Smith, A. H., Lingas, E. O., & Rahman, M. (2000). Contamination of drinking-water by arsenic in Bangladesh: a public health emergency. *Bull World Health Organ, 78*(9), 1093-1103.

UN. (2015). United Nations: The Millennium Development Goals Report 2015.

Ure, A., & Berrow, M. (1982). The elemental constituents of soils *Environmental chemistry* (pp. 94-204).

UZH. (2017). Methodenberatung *Logistische Regression*. Universität Zürich UZH.

WHO. (2011). Guidline for Drinking-water Quality *Fourth Edition*: World Health Organization WHO.

Winkel, L., Berg, M., Amini, M., Hug, S. J., & Johnson, C. A. (2008). Predicting groundwater arsenic contamination in Southeast Asia from surface parameters. *Nature Geoscience, 1*(8), 536-542. doi:10.1038/ngeo254

Yu, G., Sun, D., & Zheng, Y. (2007). Health effects of exposure to natural arsenic in groundwater and coal in China: an overview of occurrence. *Environmental health perspectives*, 636-642.

7 Index of figures and tables

Figures

Figure 1 Function of logistic regression. (UZH, 2017) 10
Figure 2 Trade-off of true negative rates (specificity) and true positive rates (sensitivity) of a logistic regression. (Analyse-IT, 2017, modified; GapWiki, 2017a) 12
Figure 3 Area under (AUC) the Receiver Operating Characteristic curve (ROC) presenting the probability of detecting true negative and true positive rates for all cut-offs. (Journals.org, 2017, modified) 13
Figure 4 Main page in GAP. 1 log in, 2 help button which leads to gapmaps.wiki, 3 toolbar, 4 display of all layers user currently works with, 5 provided GAP layers and 6 tools for showing pixel values. (GapWiki, 2017). 14
Figure 5 Prediction map of arsenic probability greater than 10 µg/l in China produced by Lado et al. Rectangles indicate area with high risk probabilities. (Rodríguez-Lado et al., 2013, modified)17
Figure 6 Distribution of arsenic measurement points. Orange points indicate the measurements data Lado et al. and this study based the calculations on, blue points are the measurement data Lado et al. used for the validation of the model. 19
Figure 7 Auxiliary variables included in this study. 20
Figure 8 Trade-off of true negative and true positive rates of the logistic regression for $model_{GAP}$ revealed in GAP. 21
Figure 9 Receiver Operating Characteristic Curve (ROC) of modelGAP and the corresponding area under the curve (AUC) retrieved in GAP 24
Figure 10 Prediction map of arsenic probability greater than 10 µg/l in China produced with GAP. Circles indicate areas with high risk probabilities. 25
Figure 11 Localised differences between the binary $model_{GAP}$ and $model_{Lado}$ (a) with the cut-off 0.41, and b) with the cut-off 0.46. 27
Figure 12 Correlations of probability $model_{GAP}$ and $model_{Lado}$ (blue) and 1-difference (%) of the binary models (yellow) as well as the corresponding AUC of different data sets of 250, 500, 750, 1000, 1250, 1500 and 2000 measurement data in threefold repetition and of the original data set of 2668 data as well as the data set of 3318 data with one repetition. 28
Figure 13 Correlations and standard deviations of probability $model_{GAP}$ and $model_{Lado}$ and binary models with 3 and 10 repetitions. 29
Figure 14 Summary of AUCs of all combinations of the four variables Holocene sediments, saline soils, subsoil texture, TWI used for $model_{GAP}$. 29

Tables

Table 1 Summary of the independent variables used in Lado et al.'s survey and frequencies and coefficients of each variable retrieved true 100 stepwise logistic regressions. (Rodríguez-Lado et al., 2013) 16
Table 2 Summary of measurement data, validation data, number and kind of variables and modelling approach in GAP and in Lado et al.'s study. 18
Table 3 Summary of data sets, repetitions and corresponding cut-offs used for the sensitivity analysis of GAP. The data set of 1250 measurement data was run in tenfold repetition to better understand the vulnerability of the correlations. 23
Table 4 Summary of all analysed combinations out of the four variables (Holocene sediments, saline soils, subsoil texture, TWI) to detect the corresponding AUC for the examination of needed variables. 23
Table 5 P-values of all variables used calculated by GAP. P-values of 0 are most probably due to rounding errors in GAP. 24
Table 6 Correlation between the probability $model_{GAP}$ and $model_{Lado}$ and congruence (%) between the binary models with the cut-off 0.41 and 0.46 for $model_{GAP}$. 26

8 Attachments

8.1 Article in GapWiki

Link to article in GapWiki

Pleas copy and past:
http://www.gapmaps.wiki/index.php?title=Modelling_Groundwater_Arsenic_Contamination_in_China_with_the_Groundwater_Assessment_Platform_(GAP)

Script of article in GapWiki

==Background==

This is the summary of the bachelor thesis of Ruth Arnheiter, student of Natural Resource Sciences at the Zurich University of Applied Sciences, ZHAW, Switzerland. Arsenic (As), occurring as a natural contaminant in groundwater, threatens the health of over 100 million people worldwide (Smedley & Kinniburgh, 2002). Prediction maps can help policymakers better identify areas at risk. In 2013, Lado et al. used logistic regression to create a prediction map of arsenic pollutants in China, basing their model on environmental variables according to the following five categories: geology, surface soil, topography, hydrology and gravity. They defined areas at high and low risk with respect to arsenic pollution by running 100 stepwise logistic regressions. Hence, the technique was not easily reproducible, especially for non-professionals. With GAP, prediction maps can be produced significantly faster and easier. But few modelling examples were available and so its sensitivity had not been determined. This bachelor thesis was focused on the demonstration and explanation of the capabilities of the platform, by (i) using GAP to produce an arsenic hazard map of China, comparable to that provided by Lado et al., (ii) analysing the differences between the two maps, and (iii) analysing the sensitivity of GAP by determining the number and kind of independent data needed to produce an effective prediction map.

==Methods==
===Modelling process in GAP===

'''Data.''' For their survey Lado et al.'s used 2668 arsenic measurements and eight environmental variables (Holocene sediments, saline soils, subsoil texture, topographic wetness index (TWI), slope, density of rivers, distance to rivers and gravity). For modelling in GAP, Lado et al.'s 2668 arsenic measurements were used as well but the latter four variables were excluded because of technical requirements of the platform, high correlations between slope and TWI, and the results of stepwise logistic regression. In addition, Lado et al. validated their model with a separate data set of 625 measurement data.

"Modelling." Logistic regression was calculated with the WHO safety guideline for arsenic in drinking water of 10 µg/L and run once. The generated prediction model (modelGAP) was transferred to ArcGIS.

'''Table 1:''' Summary of measurement data, validation data, number and kind of variables and modelling approach in GAP and in Lado et al.'s study.
{| class="wikitable"
|-
!
! "GAP"
! "Lado et al."
|-
| '''Measurement data'''
| 2688

| 2688
|-
| '''Validation data'''
| -
| 625
|-
| '''Number of variables'''
| 4
| 8
|-
| '''Variables'''
| Holocene sediments
Saline soils
Subsoil texture
TWI
| Holocene sediments
Saline soils
Subsoil texture
TWI
Slope
Density of rivers
Distance to rivers
Gravity
|-
| '''Modelling'''
| 1 Stepwise log. regression
1 logistic regression
| 100 Stepwise log. regressions
|}

===Difference analysis between the models created in GAP and in Lado et al.'s study===

'''Probability maps.''' In ArcGIS (version 10.4), modelGAP was aligned to the same grid with Lado et al.'s prediction map (modelLado) for accurate comparison. The correlation of the probabilities of modelGAP and modelLado were calculated with the band collection statistic tool in ArcGIS.

'''Binary maps.''' The models were binary coded using the cut-off values of 0.41 for modelGAP and 0.46 for modelLado. The cut-off for modelGAP was compiled in GAP by evaluating the trade-off of sensitivity (ability to correctly classify predictions above chosen threshold) and specificity (ability to correctly classify predictions below chosen threshold). The cut-off of 0.46 was determined by Lado et al. using the ROC curve (Rodríguez-Lado et al., 2013). The differences of the two binary models were located by subtracting one model from the other and the differences in percentages were calculated and compared to the prediction map. In addition, the localised differences were compared with the probability map to better understand the relation between the differences and the predicted probabilities. To estimate the influence of the cut-off modelGAP was again binary coded with the cut-off of 0.46 and processed as previously.

===Sensitivity analysis of GAP===

'''Number of measurement data.''' Once a suitable model using all 2668 measurement data was created, the effect of using fewer measurement data was tested. Randomly selected subsets of the 2668 measurement data were retrieved (250, 500, 750, 1000, 1250, 1500 and 2000 data) in threefold repetition. In addition, a data set of 3318 measurement data was aggregated from 2668 measurement data and 625 validation data used in Lado et al.'s study. There was an addition of 25 measurement data which might have been generated by the previous aggregation of the measurement points to 1 km^2 resolution. Each data set was processed according to the previous calculations with stepwise logistic regression to calculate the relative and binary correlations using the corresponding cut-offs compiled in GAP. In addition, for a better understanding of the

vulnerability of the correlations, the model with 1250 measurement data was recalculated in tenfold repetition.

'''Number and kind of variable.''' To examine the effect of using fewer variables on the modelling, the validity of all combinations of the four variables used in ModelGAP (Holocene sediments, saline soils, subsoil texture, TWI) were analysed by comparing the corresponding AUC. Every combination of variables was joined with the 2668 measurement data to run a logistic regression. To prove that non-related variables produce an AUC around 0.5, the 2668 measurement data was joined with a random variable temperature to calculate the corresponding AUC.

==Results==
===Probability model created in GAP===
In modelGAP, all variables showed p-values < 0.01, except for soil texture which presented a p-value of 0.03 (Table 2). Highest coefficient showed saline soils (0.88), followed by Holocene sediments (0.73), soil texture (-0.25) and TWI (0.18) (Table 2). Besides, soil texture showed a negative coefficient. The AUC of the model was 0.68. ModelGAP showed probabilities between 20 - 80% (Figure 3). High risk areas were determined in the Xinjiang and Quinghai province, Hetao-Huhhot and Liao-Ho basin.
'''Table 2:''' P-values of all variables used calculated by GAP. P-values of 0 are most probably due to rounding errors in GAP.
{| class="wikitable"
|-
! '''Variable'''
! "p-value"
! "coefficients"
|-
| '''Holocene sediments'''
| 0
| 0.73
|-
| '''Saline soils'''
| 0
| 0.88
|-
| '''Soil texture''' (clayey soils)
| 0.031
| -0.25
|-
| '''TWI'''
| 0
| 0.18
|}
===Analysed differences between modelGAP and modelLado===

The probabilities of modelGAP and modelLado were highly correlated (0.97) (Table 3). The binary model using the cut-off of 0.41 for modelGAP differed in 6% of the areas (Figure 4). 95% of these differences lied in areas with predicted probabilities of high arsenic contamination in groundwater between 40 – 50%. Using the cut-off of 0.46 for modelGAP, the two binary modelGAP and modelLado were brought more in line. The differences became smaller; in fact, as low as 1.5%.

'''Table 3:''' Correlation between the probability modelGAP and modelLado and congruence (%) between the binary models with the cut-off 0.41 and 0.46 for modelGAP.
{| class="wikitable"
|-

!
! "Correlation"
"ModelGAP"

"(probability)"
! "Congruence (%)"
“ModelGAP”
“(binary; cut-off 0.41”
! "Congruence (%)"
“ModelGAP”
“(binary; cut-off0.46)”
|-
| ModelLado>
(probability)
| 0.97
| -
| -
|-
| ModelLado
(binary; cut-off 0.46)
| -
| 6
| 1.5
|}

===Analysed sensitivity of GAP===

As expected, the best result was achieved with 2668 measurement data. Most of the randomly selected subsets above the number of 1000 measurement data showed similar results compared to the initial 2668 measurement data. The probability models were well correlated (0.95), the binary models were congruent at 90% and the AUC was not smaller than 0.67 for any data set (Figure 5). Data sets lower than 1000 measurement data achieved less reliable results, considering means, variations and AUCs. In addition, the data set of 1250 measurement data showed great variations, especially with the binary models. This variation, however, decreased if the calculations were repeated multiple times. Testing the number of required variables showed that the best result was achieved with the initial four variables, namely Holocene sediments, saline soils, subsoil texture, and TWI (0.68) (Figure 6). The smallest AUC was shown with the variable subsoil texture (0.5). An AUC of 0.65 was achieved by using either three variables (Holocene sediments, subsoil texture and saline soils) or two variables (Holocene sediments and TWI or Holocene sediments and saline soils), while temperature showed an AUC of 0.51, indicating that temperature does not contribute to arsenic contamination in groundwater.

==Discussion==

GAP predicted the same areas to be high risk areas as Lado et al. Holocene sediments, saline soils, subsoil texture and TWI are indeed the more important variables, compared to slope, density of rivers, distance to rivers and gravity, as used in Lado et al.’s survey (Rodríguez-Lado et al., 2013). The inverse relation of subsoil texture showed that the presence of clayey soils decreases the probability of arsenic in groundwater. The best achieved AUC in GAP was 0.68, which is considered to be too low for a good prediction model (Hosmer, Lemeshow, & Sturdivant, 2013). Nevertheless, GAP produces results as accurate as those of Lado et al. by running only one logistic regression, compared to the 100 stepwise logistic regressions that Lado et al.’s used. Despite the different number of variables (four variables for modelGAP, eight variables for modelLado), the two models’ probabilities were highly correlated (0.97). The binary model with the cut-off of 0.41 slightly differed from the binary model of Lado et al. (6%). Interestingly, almost all localised differences were in areas with risk probabilities between 40 - 50% (95%). Approximating the cut-off for modelGAP to that of modelLado the area of differences decreased to one fourth of

the differences revealed with the smaller cut-off (1.5%). Hence, the threshold used for the binary coding is decisive about the correlation of two binary models. Nevertheless, whatever technique was used for the analysis of differences, the two models showed a congruency of 94%, which is high. Most of the randomly selected subsets of the 2668 measurement data showed a similar behaviour compared to the initial data set. As expected, the bigger the data set was, the better the models were correlated. Interestingly, similar results (compared to the 2668 measurement data) were achieved with as little as 1000 measurement data, which is less than 40% of the initial data set. However, these investigations were based on only three repetitions which may not be enough to be conclusive, especial when standard deviations vary, such as for 1250 measurement data. But the investigations with ten repetitions proved that the higher the number of repetitions, the smaller was the variation. The investigation of the variables showed the best results with four variables. Interestingly, in some cases the number of variables did not determine the AUC but rather the composition of the variables. However, all of the combinations out of the four variables produced a smaller AUC than the four variables themselves. Temperature as a variable showed that an AUC of 0.51 indicates non-correlation. Nevertheless, this study does not give final recommendations regarding the number of required measurement data in relation to the area of investigation. For example, downscaling these results to Switzerland, which has a surface area of 41,285 km^2 (which is less than 0.5 % of China's 9,297,000 km^2), would mean that less than 0.5% of 1000 measurement data should be sufficient to predict arsenic pollution in Switzerland, which intuitively is insufficient. In the future, investigation on the smallest measurement data set related to the area but still produces reliable results will help policymakers better determine how many measurement points are required. In addition, this survey did not analyse the potential influence of the local distribution of measurement data. Investigating whether such distribution is indeed of importance for the generation of prediction maps should be of considerable interest for further studies.

==Conclusion and Outlook==

By running only one logistic regression, GAP produces results as accurate as those of Lado et al. who run 100 logistic regressions. The data could even be limited to a certain number of measurement data (1000 measurement data in this survey) as well as to a certain selection of variables (four variables in this survey). The selection of the most important variables was essential though. Therefore, GAP should (i) allow users to add any number of variables and (ii) give information on how GAP defines a variable to fit or not to fit. These improvements of GAP will help users to make the right selection of variables and to better understand the techniques behind the platform, which will enhance GAP's promotion and dissemination.

The following further investigation could be of interest:

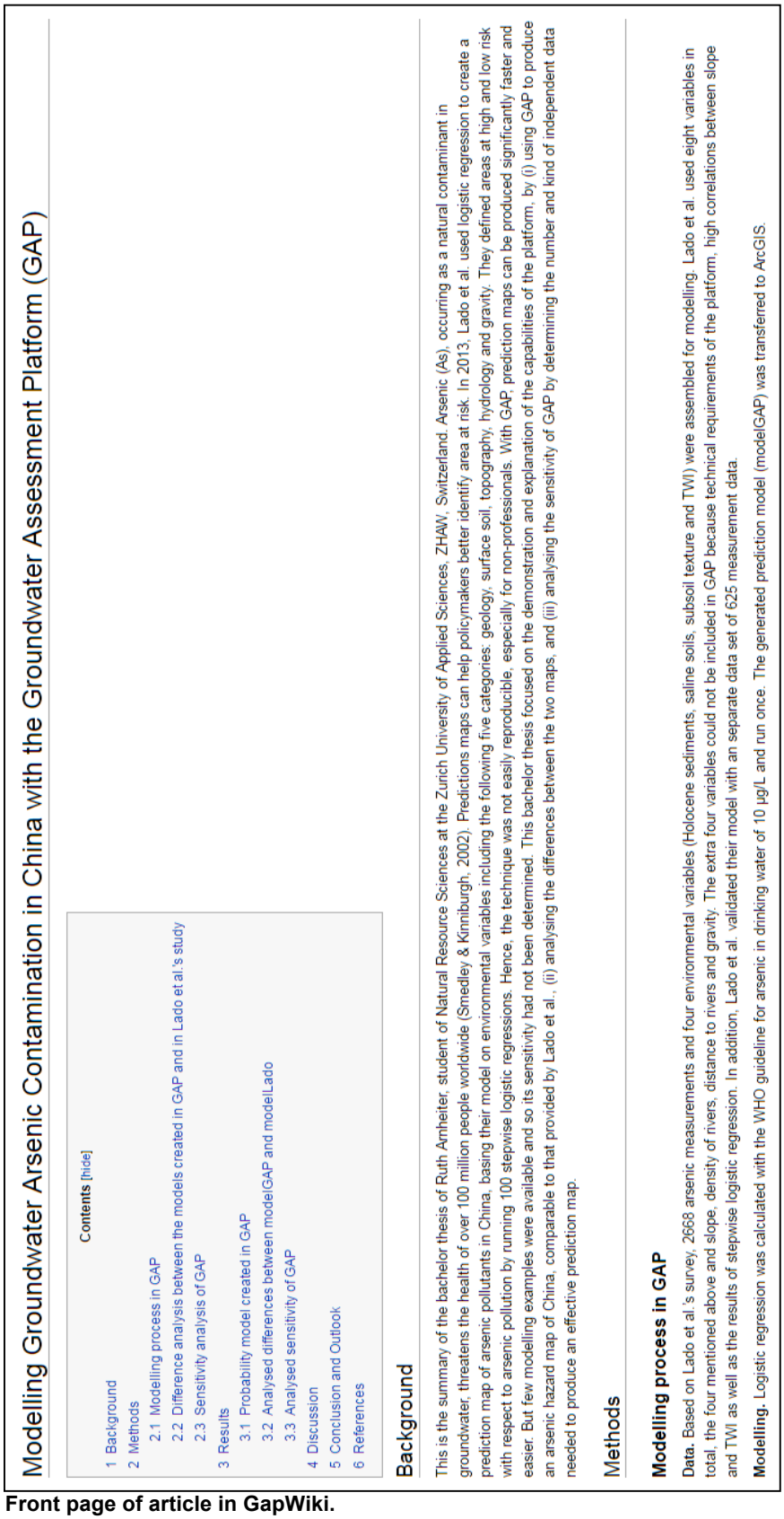
*Which is the smallest measurement data set related to the area of interest but still produces reliable results?
**This will help policymakers to determine the number of measurements points required.

*How does the distribution of measurement data affect the probability model?
**This will help policymakers to determine the distribution of measurements points required.

==References==

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. (2013). "Applied Logistic Regression, Third Edition:" Willey, New Jersey.
Rodríguez-Lado, L., Sun, G., Berg, M., Zhang, Q., Xue, H., Zheng, Q., & Johnson, C. A. (2013). Groundwater arsenic contamination throughout China. "Science, 341"(6148), 866-868.
Smedley, P. L., & Kinniburgh, D. G. (2002). A review of the source, behaviour and distribution of arsenic in natural waters. "Applied Geochemistry, 17"(5), 517-568. doi:Pii S0883-2927(02)00018-5 Doi 10.1016/S0883-2927(02)00018-5

Front page of article in GapWiki



Modelling Groundwater Arsenic Contamination in China with the Groundwater Assessment Platform (GAP)

Contents [hide]

1 Background
2 Methods
2.1 Modelling process in GAP
2.2 Difference analysis between the models created in GAP and in Lado et al.'s study
2.3 Sensitivity analysis of GAP
3 Results
3.1 Probability model created in GAP
3.2 Analysed differences between modelGAP and modelLado
3.3 Analysed sensitivity of GAP
4 Discussion
5 Conclusion and Outlook
6 References

Background

This is the summary of the bachelor thesis of Ruth Amheiter, student of Natural Resource Sciences at the Zurich University of Applied Sciences, ZHAW, Switzerland. Arsenic (As), occurring as a natural contaminant in groundwater, threatens the health of over 100 million people worldwide (Smedley & Kinniburgh, 2002). Predictions maps can help policymakers better identify area at risk. In 2013, Lado et al. used logistic regression to create a prediction map of arsenic pollutants in China, basing their model on environmental variables including the following five categories: geology, surface soil, topography, hydrology and gravity. They defined areas at high and low risk with respect to arsenic pollution by running 100 stepwise logistic regressions. Hence, the technique was not easily reproducible, especially for non-professionals. With GAP, prediction maps can be produced significantly faster and easier. But few modelling examples were available and so its sensitivity had not been determined. This bachelor thesis focused on the demonstration and explanation of the capabilities of the platform, by (i) using GAP to produce an arsenic hazard map of China, comparable to that provided by Lado et al., (ii) analysing the differences between the two maps, and (iii) analysing the sensitivity of GAP by determining the number and kind of independent data needed to produce an effective prediction map.

Methods

Modelling process in GAP

Data. Based on Lado et al.'s survey, 2668 arsenic measurements and four environmental variables (Holocene sediments, saline soils, subsoil texture and TWI) were assembled for modelling. Lado et al. used eight variables in total, the four mentioned above and slope, density of rivers, distance to rivers and gravity. The extra four variables could not be included in GAP because technical requirements of the platform, high correlations between slope and TWI as well as the results of stepwise logistic regression. In addition, Lado et al. validated their model with an separate data set of 625 measurement data.

Modelling. Logistic regression was calculated with the WHO guideline for arsenic in drinking water of 10 µg/L and run once. The generated prediction model (modelGAP) was transferred to ArcGIS.

Front page of article in GapWiki.

8.2 Signature plagiarism (German)

Zürcher Hochschule für Angewandte Wissenschaften

zh aw Life Sciences und Facility Management

Erklärung betreffend das selbständige Verfassen einer Bachelorarbeit im Departement Life Sciences und Facility Management

Mit der Abgabe dieser Bachelorarbeit versichert der/die Studierende, dass er/sie die Arbeit selbständig und ohne fremde Hilfe verfasst hat.

Der/die unterzeichnende Studierende erklärt, dass alle verwendeten Quellen (auch Internetseiten) im Text oder Anhang korrekt ausgewiesen sind, d.h. dass die Bachelorarbeit keine Plagiate enthält, also keine Teile, die teilweise oder vollständig aus einem fremden Text oder einer fremden Arbeit unter Vorgabe der eigenen Urheberschaft bzw. ohne Quellenangabe übernommen worden sind.

Bei Verfehlungen aller Art treten Paragraph 39 und Paragraph 40 der Rahmenprüfungsordnung für die Bachelor- und Masterstudiengänge an der Zürcher Hochschule für Angewandte Wissenschaften vom 29. Januar 2008 sowie die Bestimmungen der Disziplinarmassnahmen der Hochschulordnung in Kraft.

Ort, Datum: Winterthur, 2.6.2017

Unterschrift: .....